

*HMM's –
Hidden Markov Models
Part I*

Was ist ein „hidden Markov model“ (HMM)?

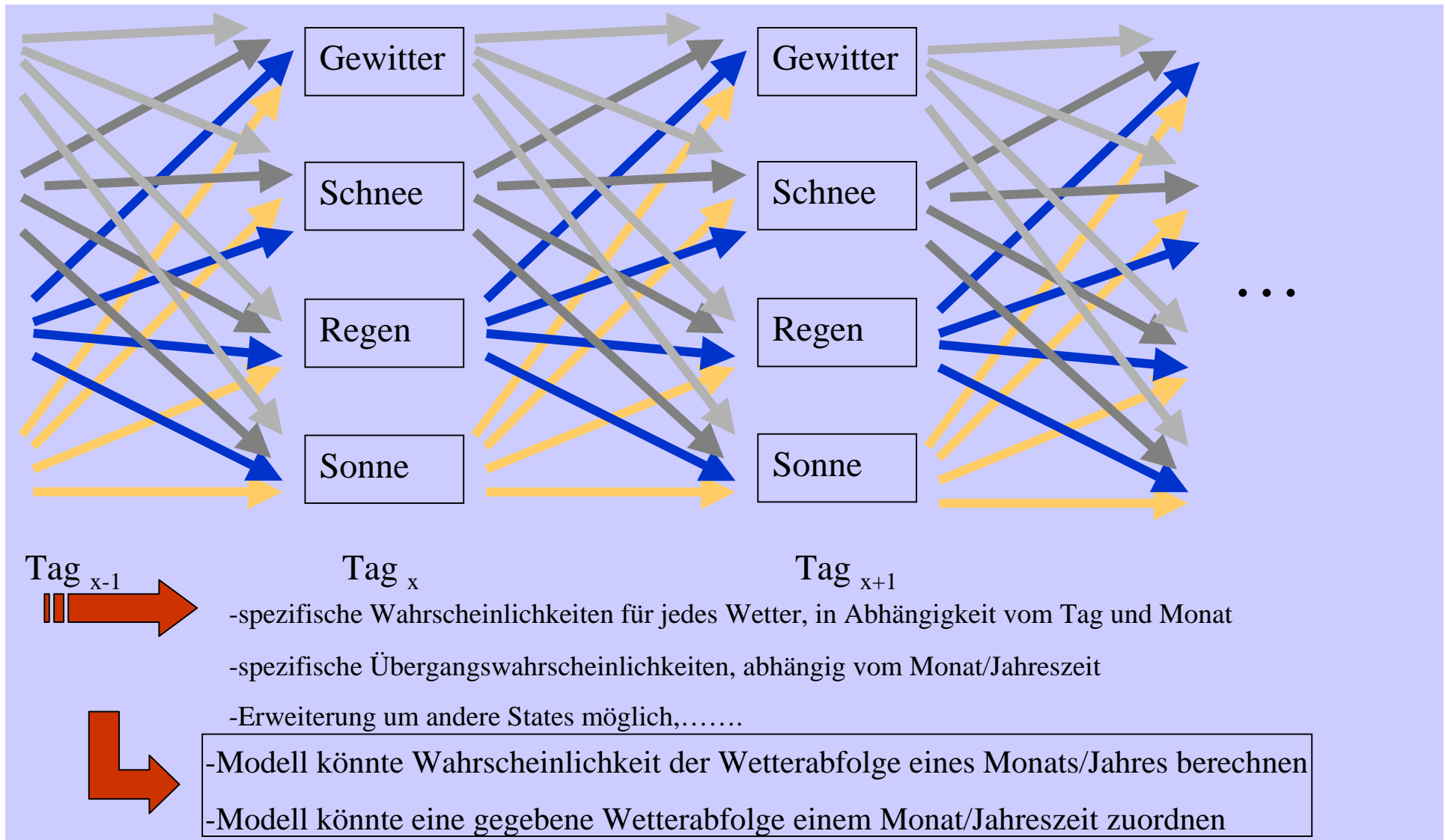
Markov-Kette: - Serie von unabhängigen Ereignissen

Bsp: In einer Kette der Ereignisse **A-B-C-D** ist die Wahrscheinlichkeit, dass auf Beobachtung **C** die Beobachtung **D** folgt, nur abhängig von Ereignis **C**, nicht aber von **A** und **B**

Markov-Modell:

- HMM ist ein endliches Modell,
- beschreibt eine Reihe von Beobachtungen durch einen versteckten stochastischen Prozess

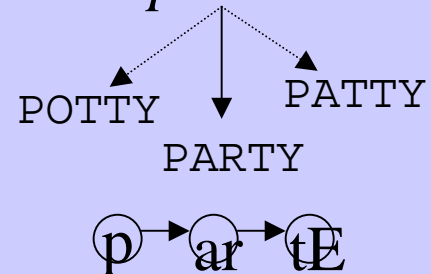
Bsp. für ein potentielles HMM-Modell: **Wettervorhersage**



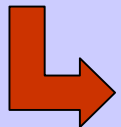
Bsp: HMMs in Spracherkennung

- **Sprache:** Kette von Lauten
- **Beobachtungen** = Laute, die ein Wort bilden

“lets ‘gO tu the ‘pär-tE”



Jede Soundsequenz kann mit einer bestimmten Wahrscheinlichkeit von einem Modell „generiert“ werden.



Das Modell wird allen Soundsequenzen, die Lauten des modellierten Wortes ähneln, hohe Wahrscheinlichkeiten zuweisen.

Anwendungen von HMMs

Spracherkennung:

Alphabet = Phoneme, aus denen Worte konstruiert werden

Beobachtungen = Ketten von Lauten

Sequenz-Alignment/Proteinmodellierung:

20 Aminosäuren = „Alphabet“, bauen Proteine auf

Beobachtungen = Ketten von AS (Primärstruktur)

Regular expressions

- A regular expression is a pattern that can match various text strings
- C. elegans vs. Caenorhabditis elegans

`C[\.a-z]* elegans`

For syntax of regular expressions see

http://www.cs.utah.edu/dept/old/texinfo/emacs19/emacs_17.html#SEC83

Example of coding sequences by regular expressions

A	C	A	-	-	-	A	T	G
T	C	A	A	C	T	A	T	C
A	C	A	C	-	-	A	G	C
A	G	A	-	-	-	A	T	C
A	C	C	G	-	-	A	T	C

`[AT][CG][AC][ACGT]*A[ATG][GC]`

Problem with this regular expressions

A	C	A	-	-	-	A	T	G
T	C	A	A	C	T	A	T	C
A	C	A	C	-	-	A	G	C
A	G	A	-	-	-	A	T	C
A	C	C	G	-	-	A	T	C

TGCT--AGG (very implausible)

and

ACAC--ATC (consensus sequence)

can be both derived from

$[AT][CG][AC][ACGT]^*A[ATG][GC]$

HMM derived from this alignment

A	C	A	-	-	-	A	T	G
T	C	A	A	C	T	A	T	C
A	C	A	C	-	-	A	G	C
A	G	A	-	-	-	A	T	C
A	C	C	G	-	-	A	T	C

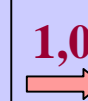
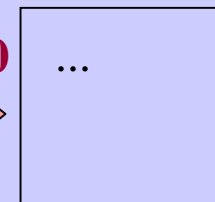
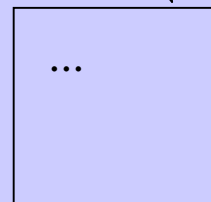
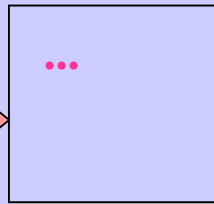
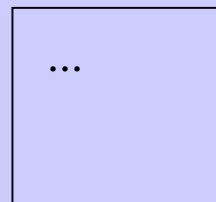


0,4

A	0,2
C	0,4
G	0,2
T	0,2

5 transitions in total from the insert state, the probability of making a transition to itself is 2, the probability of transition back into match state is 3 (all three sequences with insertions)

A	0,8
C	0
G	0
T	0,2

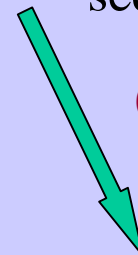


A	0
C	0,8
G	0,2
T	0

0,6



0,6



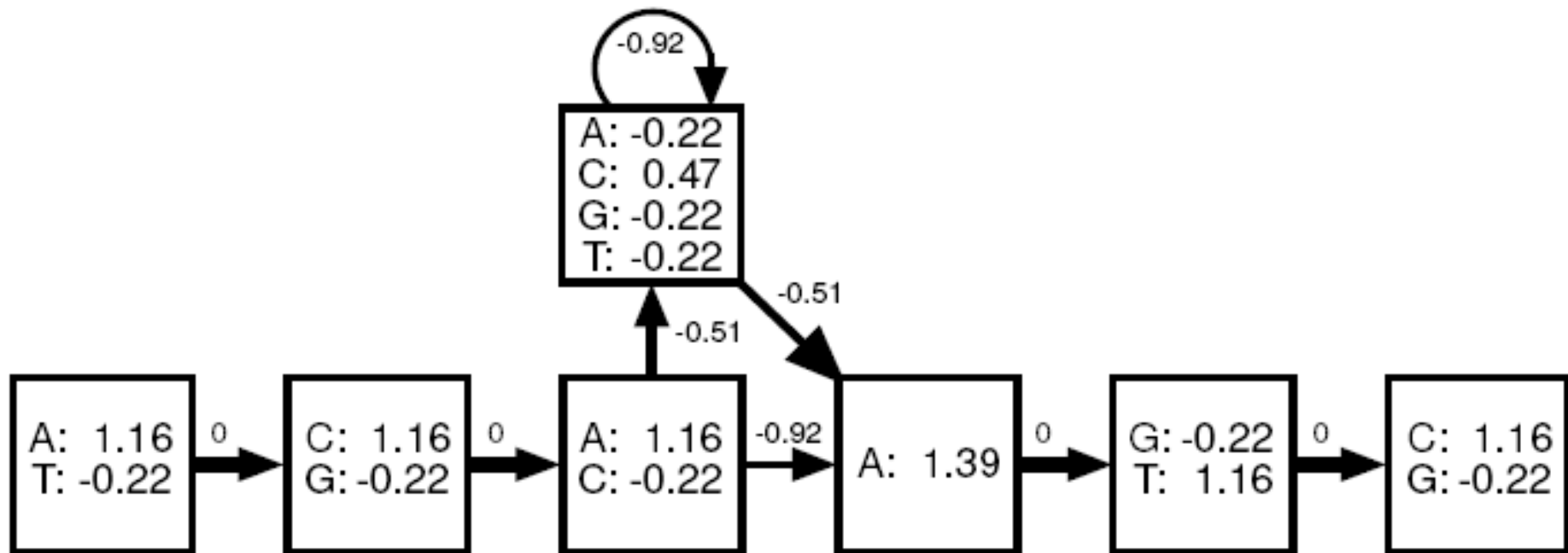
Probabilities and log-odds scores

	Sequence	Probability x 100	Log odds
consensus	ACAC--ATC	4.7	6.7
original	ACA---ATG	3.3	4.9
sequences	TCAACTATC	0.0075	3.0
	ACAC—AGC	1.2	5.3
...	AGA---ATC	3.3	4.9
exceptional	TGCT--AGG	0.0023	-0.97

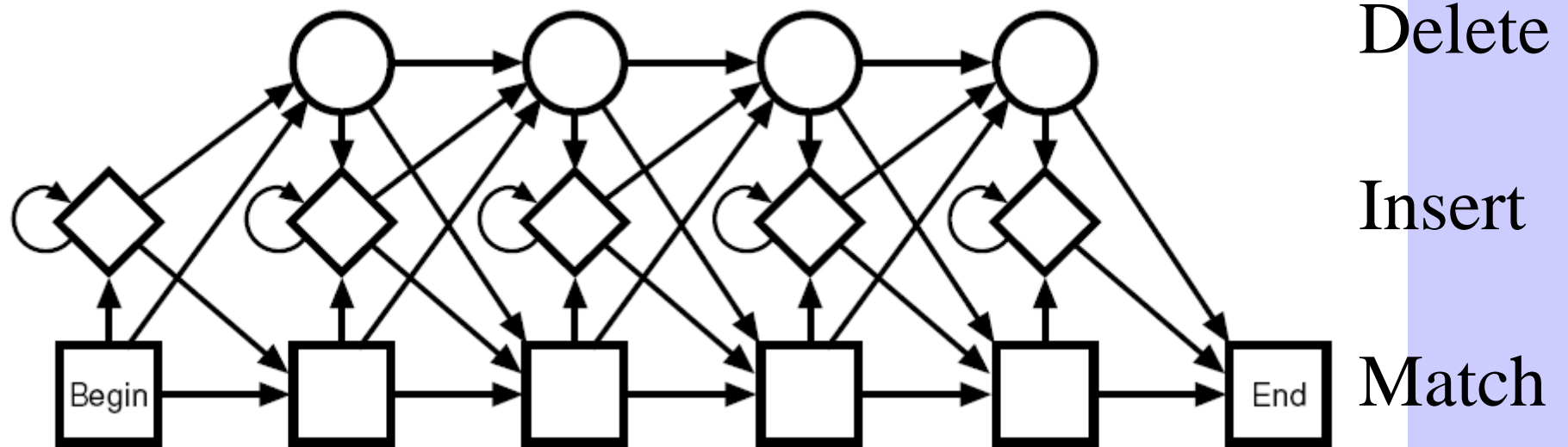
$$\begin{aligned}
 P(\text{ACACATC}) &= 0.8 \times 1 \times 0.8 \times 1 \times 0.8 \times 0.6 \times \\
 &\quad 0.4 \times 0.6 \times 1 \times 1 \times 0.8 \times 1 \times 0.8 \\
 &\simeq 4.7 \times 10^{-2}.
 \end{aligned}$$

$$\text{log-odds for sequence } S = \log \frac{P(S)}{0.25^L} = \log P(S) - L \log 0.25.$$

Log-odds scores representation of HMM



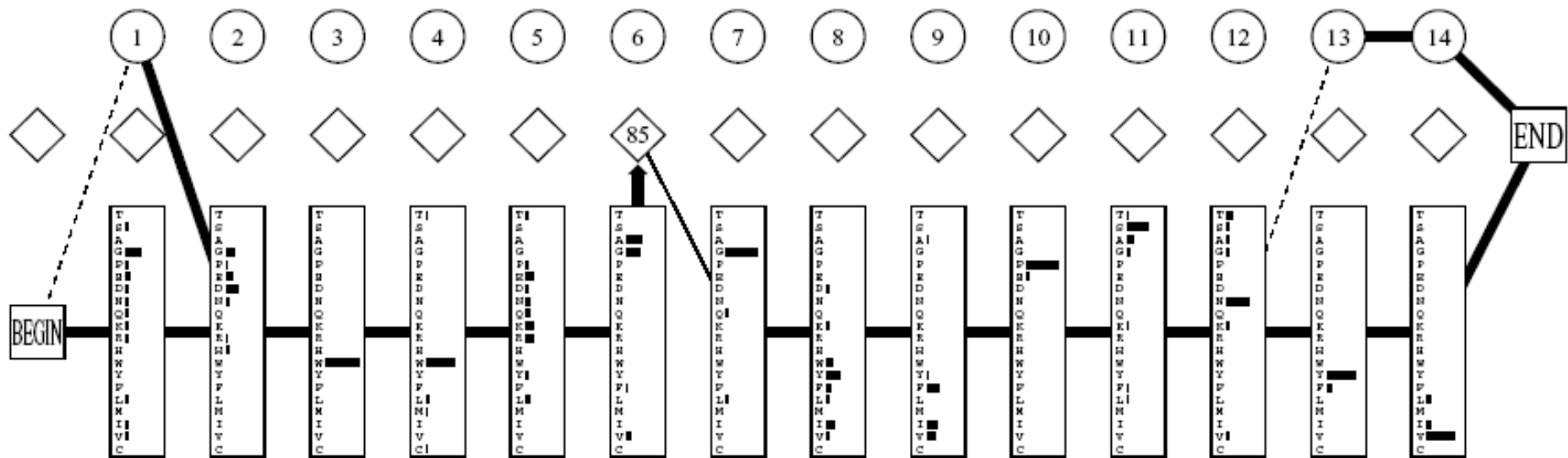
Structure of profile HMM



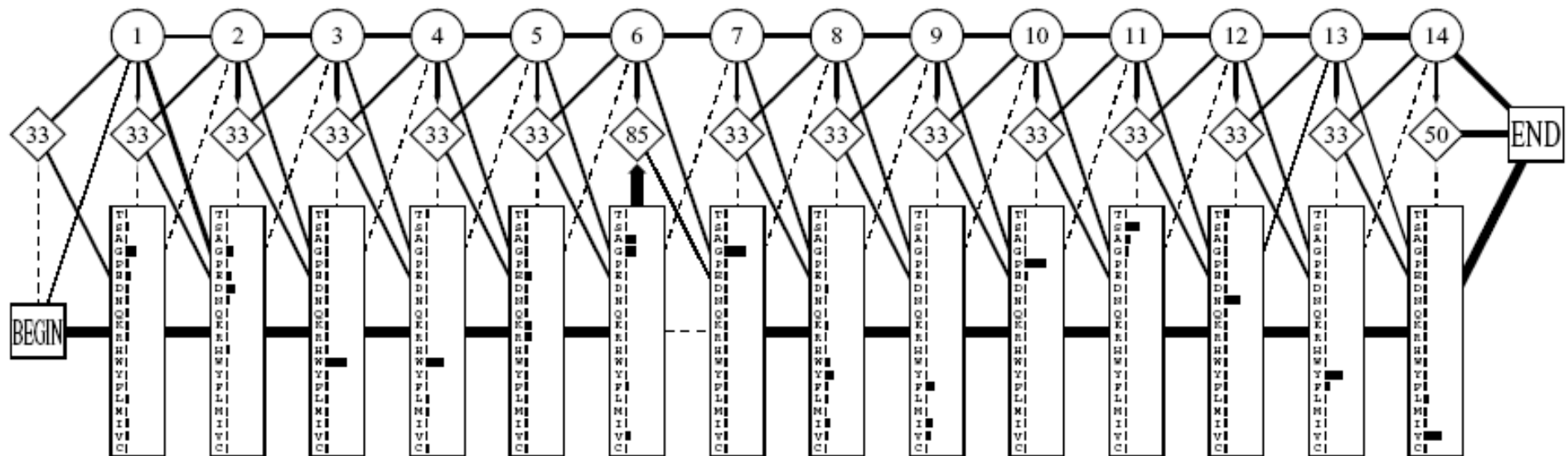
Alignment of 30 proteins of SH3 domain

```
GGWWRG d y . g g k k q L W F P S N Y V
IGWLN G y n e t t g e r L G D F P G T Y V
PNWWE G q l . . n n r r G I F P S N Y V
DEWWE Q A r r . . d e q i G I V P S K - -
GEWWE K A q s . . t g q e G F I P F N F V
GDWWE L A r s . . s g q t G Y I P S N Y V
GDWWE D A e l . . k g r r G K V P S N Y L
- D W W E A r s l i s s g h r G Y V P S N Y V
GDWWE Y A r s l i t n s e G Y I P S T Y V
GEWWE K A r s l a t r k e G Y I P S N Y V
GDWWE L A r s l v t g r e G Y V P S N F V
GEWWE K A k s l s k r e G F I P S N Y V
GEWWE C E A q t . k n g q . G W V P S N Y I
SDWWE R V v n l t t r q e G L I P L N F V
LPWWE R A r d . k n g q e G Y I P S N Y I
RDWWE E F r s k t v y t p G Y Y E S G Y V
EHWWE K V k d . a l g n v G Y I P S N Y V
IHWWE R V q d . r n g h e G Y V P S S Y L
KDWWE K V e v . n d r q G F V P A A Y V
VGWWE M P G l n e r t r q r G D F P G T Y V
PDWWE E G e l . . n g q r G V F P A S Y V
ENWWE N G e i . . g n r k G I F P A T Y V
EEWWE L E G e c . . k g k v G I F P K V F V
GGWWE K G d y . g t r i q Q Y F P S N Y V
DGWWE R G s y . . n g q v G W F P S N Y V
QGWWE R G e i . . y g r v G W F P A N Y V
GRWWE K A r r . a n g e t G I I P S N Y V
GGWWE T Q G e l . k s g q k G W A P T N Y L
GDWWE E A r s n . t g g e n G Y I P S N Y V
NDWWE T G r t . . n g k e G I F P A N Y V
```

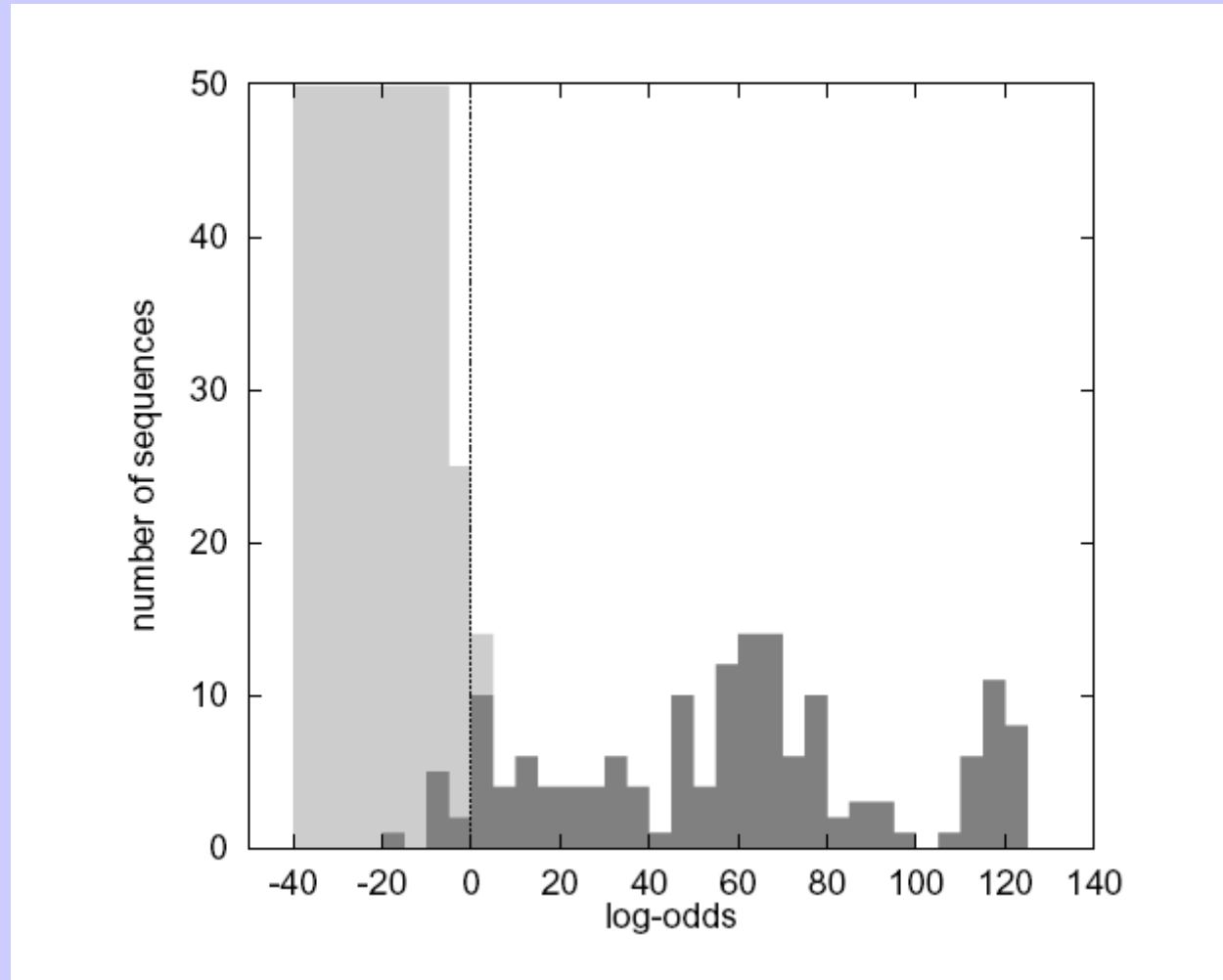
HMM model of alignment of 30 proteins of SH3 domain



HMM model of alignment of 30 proteins of SH3 domain using pseudocounts of 1



Searching SwissProt with HMM profile of SH3 domain



Application areas for HMM's:

Does this sequence belong to a particular family?

Can we identify regions in a sequence (for instance –
alpha helices, beta sheets)?

...and less directly:

Pairwise/multiple sequence alignment

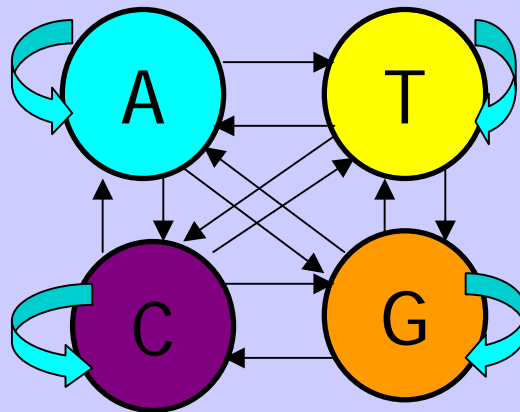
Searching databases for protein families (building
profiles).

Markov Chains (reminder)

A sequence of random variables X_1, X_2, \dots where each present state depends only on the previous state.

$A \rightarrow C \rightarrow G \rightarrow G \rightarrow T \rightarrow A \dots$ (vertical or horizontal!)

These conditional probabilities can be illustrated as follows (in DNA):



Markov chain probabilities

Each arrow has a transition probability

$$P_{CA} = P(x_i=A|X_{i-1}=C)$$

Thus – the probability of a sequence x will be :

$$P(x) = P(x_L, x_{L-1}, \dots, x_1) = P(x_1) \cdot \prod_{i=1}^L P_{x_i - 1 x_i}$$

Some terminology...

We differentiate between **states** and **symbols**:

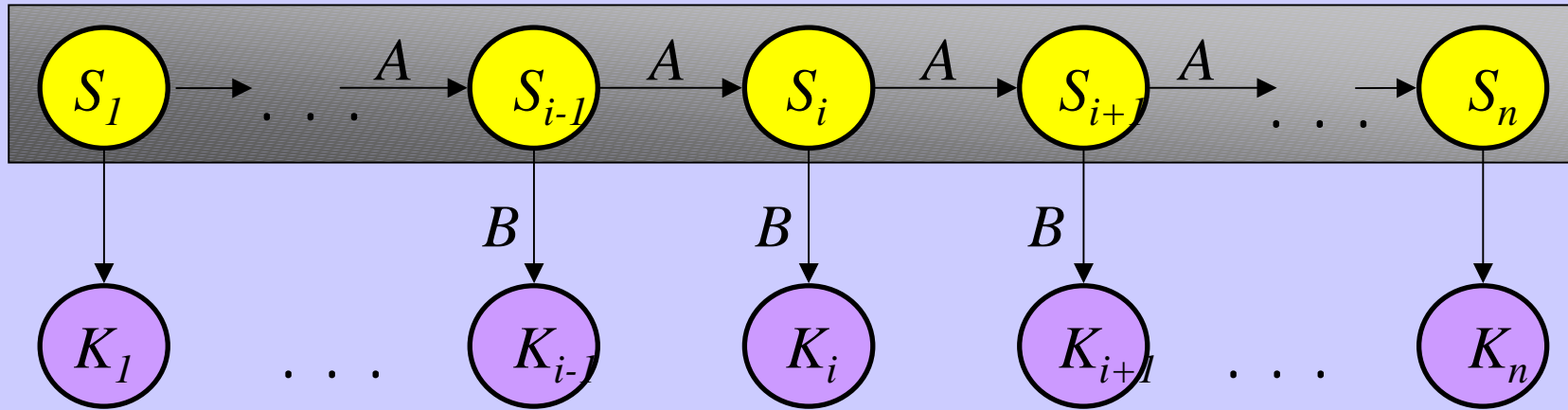
Symbols are what we see, our observations – the sequence we see.

States are what we talk of in theory. They are part of the probabilistic model

(P_{st} – the probability of the transition between 2 theoretic states, s and t. Here s and t are A , C, G or T).

We will see the importance of this difference later on.

HMM Formalism



$\{S, K, P, A, B\}$

$S: \{s_1 \dots s_N\}$ are the values for the hidden states

$K: \{k_1 \dots k_M\}$ are the values for the observations

The hidden states **emit/generate** the symbols (observations)

$\Pi = \{\pi_i\}$ are the initial state probabilities

$P = \{P_{ij}\}$ are the state transition probabilities

$B = \{b_{ik}\}$ are the emission probabilities (*which in our case are 0 or 1:*

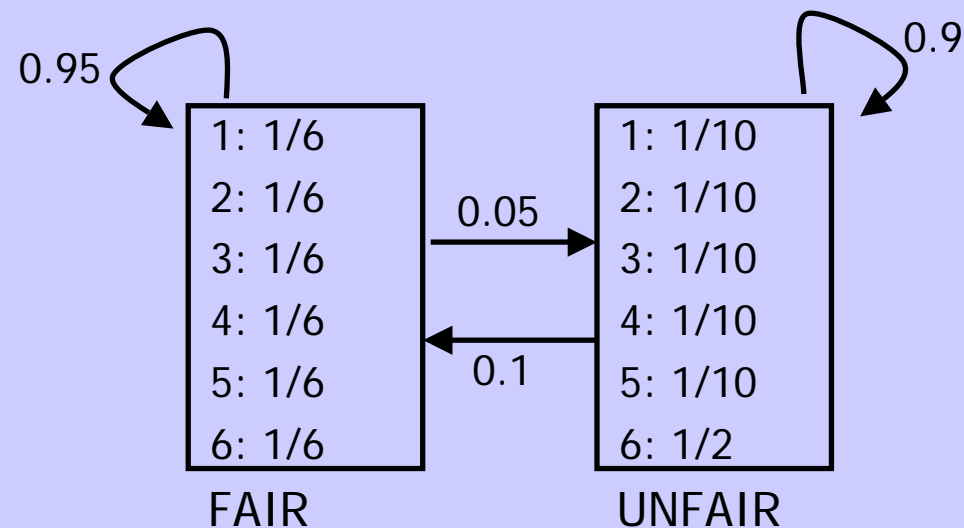
$P(A/A+) = P(A/A-) = 1, P(A/C+) = 0$)

Markov chain vs. HMM

The state sequence itself follows a simple Markov chain. But-
The essential difference between a Markov chain and an HMM is it is no longer possible to know the state by looking at the symbols – the state is **hidden**.

Another example – the dishonest casino

In a casino, they use a fair die most of the time, but occasionally switch to an unfair die. The **switch between dice** can be represented by an HMM:



Dishonest casino - continued

The symbols (observations) are the sequence of rolls:

3 5 6 2 1 4 6 3 6...

What is hidden?

If the die is fair or unfair:

f f f f u u u f f

This is a Markov chain. Except for that, we have:

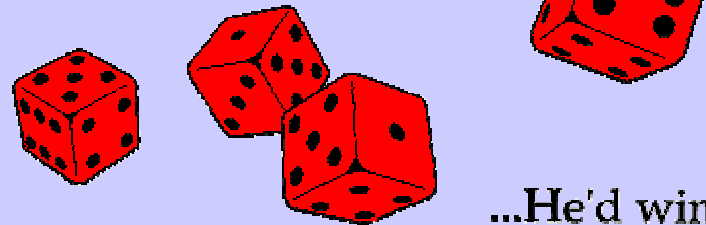
Emission probabilities:

Given a state, we have 6 possible matching symbols, each with an emission probability.

Exposing the casino

Once again – we can estimate which is the most probable state path, and estimate when the casino was cheating

If God played dice...



...He'd win